Comparative analyses of spectrum and pitch modeling by HMMs applied to a Brazilian Portuguese TTS

F. L. de F. BARBOSA[2], G. de O. PINTO[2], R. da S. MAIA[1], F. G. V. RESENDE Jr.[2]

PEE/COPPE[1], DEL/EE[2]
Universidade Federal do Rio de Janeiro
PO Box 68504, 21945-970, Rio de Janeiro, RJ
BRAZIL
{filipe, guilherme, maia, gil}@lps.ufrj.br

Abstract: - This work presents a Brazilian Portuguese TTS based on HMM modeling of spectrum and pitch. A major advantage of HMM-based TTS systems is the possibility of fast speaker adaptation. During the HMM training, mel-cepstral coefficients are used to represent the vocal tract and pitch information is obtained through the autocorrelation method. Comparative analysis show the effectiveness of the proposed HMM-based TTS, as well as the impact of HMM modeling for spectrum and pitch.

Key-words: - Speech Synthesis, Spectrum, Pitch, State Duration, HMM, Brazilian Portuguese TTS

1 Introduction

Text-to-speech (TTS) synthesis is an area which has been facing growing interests in the research and commercial levels. The main target of researchers is to increase the naturalness through a better understanding and modeling of prosody features.

Recently, TTS systems based on hidden Markov model (HMM) have been introduced [1, 2]. The usage of HMM to model the spectrum offers an important advantage so-called speaker adaptation. It has been shown that only ten sentences are sufficient to adapt HMMs from one speaker to another [3].

In this work, HMMs are used to model not only spectrum but also pitch information. To generate the present HMM TTS for the Brazilian Portuguese language, a group of 200 phonetically balanced sentences was recorded and processed. Mel-cepstral coefficients which can represent speech spectrum were extracted [4], pitch information was obtained through the autocorrelation method, and HMM state durations were determined from the transition matrices for each model. Experiments have been undertaken in such a way that the effect of using HMM to model spectrum and pitch, separately and combined, could be examined.

This paper is organized as follows. In Section 2, a detailed description of the proposed HMM TTS is introduced; Section 3 shows experiment results; and conclusions are presented in Section 4.

2 The TTS synthesizer

Figure 1 illustrates the HMM TTS synthesis procedure. First, the text to be synthesized is transcribed by a context dependent phone level transcription software [5, 6]. It can be noticed that a trained HMM database is the main requirement from which speech can be produced. In the following, each part of the proposed TTS is described separately.

2.1 Database HMMs

2.1.1 Spectrum and pitch models

To train the HMMs, we used a speech database composed of 200 phonetically balanced unlabeled utterances from one speaker. Speech was

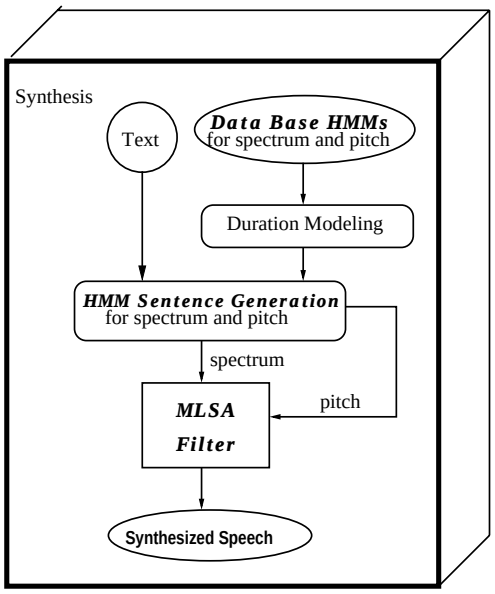

Figure 1: Illustration of the synthesis procedure for the proposed TTS.

sampled at 16 kHz.

For each utterance on the speech database, mel-cepstral coefficients and pitch were obtained using 25 ms Hamming windows with 5 ms shifts. These parameters were used to train 5-state left-to-right HMMs with no skips. For pitch modeling, the feature vector was composed of one single coefficient while for the spectrum modeling 25 coefficients were used (including the 0^{th} order one).

In our general purpose TTS we use the flat-start initialization. In the training procedure a global mean and variance is calculated and set all the Gaussian distributions for every monophone HMM, then the Baum-Welch embedded training re-estimation algorithm was performed. After that, monophones HMM were cloned, generating, in this way, HMM triphones. The new models were re-estimated with the embedded training.

Also, we perform the bootstrap initialization of some sentences separately, since we want to know the HMM TTS upper quality limit of spectrum and pitch modeling under our constraints.

2.1.2 Spectrum and pitch extraction

Speech spectrum $H(e^{jw})$ can be obtained from $M+1$ mel-cepstral coefficients $\{c_0,\ldots,c_M\}$ through [4]

$$H(z) = \exp \sum_{m=0}^{M} c_m \tilde{z}^{-m}, \qquad (1)$$

under the constraint

$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}, \quad |\alpha| < 1. \qquad (2)$$

The phase is modeled by the factor α. For a sampling rate of 16kHz, the choice $\alpha = 0.42$ gives a good approximation to the mel frequency scale. Since the transfer function $H(z)$ can be written as

$$H(z) = K \cdot D(z), \qquad (3)$$

where K is a constant, the mel-cepstral coefficients are calculated by

$$\epsilon = \min_{D(z)} E[e^2(n)], \qquad (4)$$

where $e(n)$ is the output of the inverse filter $\frac{1}{D(z)}$ when the input is a single pulse. For the present TTS, we use $M = 24$.

The pitch analysis is performed using the biased autocorrelation method, where the pitch lag value is $\tau = 20,\ldots,160$ which maximizes the autocorrelation $r(\tau)$, given by

$$r(\tau) = \sum_{n=0}^{N-1} s(n)s(n-\tau), \qquad (5)$$

with $s(n)$ being the windowed input speech frame and N its related window length.

2.1.3 State duration estimation

The state durations can be estimated from the models according to [7]

$$d_{q_j} = \frac{1}{1-a_{q_j,q_j}}, \qquad (6)$$

where d_{q_j} and a_{q_j,q_j} are the estimated duration and the diagonal transition matrix element, respectively, for the state q_j from a given HMM model.

2.2 HMM sentence generation

2.2.1 Parameter Generation

From an HMM λ, with a given state sequence $\mathbf{Q} = \{q_1, q_2, \ldots, q_T\}$, the output observation vector sequence $\mathbf{O} = \{\vec{o}_1, \vec{o}_2, \vec{o}_3, ..., \vec{o}_T\}$ is extracted in such a way that [8, 9]

$$\log P[\mathbf{O}|\mathbf{Q}, \lambda] = -\frac{1}{2}\sum_{j=1}^{T}\left((\vec{o}_j - \vec{\mu}_{q_j})^t \mathbf{U}_{q_j}^{-1}(\vec{o}_j - \vec{\mu}_{q_j})\right)$$

$$-\frac{1}{2}\sum_{j=1}^{T}\log|\mathbf{U}_{q_j}| - \frac{M'T}{2}\log 2\pi \qquad (7)$$

is maximized when the speech parameter vector sequence $\mathbf{O} = \{\vec{o}_1, ..., \vec{o}_T\}$ is equal to the mean vectors, i.e.,

$$\vec{o}_j = \vec{\mu}_{q_j}, \qquad 1 \leq j \leq T, \qquad (8)$$

where $\vec{\mu}_{q_j}$ and $\mathbf{U}_{q_j}$ are the mean vector and the covariance matrix, respectively, for the state q_j, and M' represents the length of the mean vector. Spectrum models use $M' = 25$ while pitch models use $M' = 1$ for the present TTS.

2.2.2 Parameter sequences

Since only the static parameters of the mel-cepstral coefficients and pitch are considered, the parameter sequences will be formed by concatenating the mean vectors $\vec{\mu}_{q_j}$ of every state q_j in each HMM that belongs to the HMM sequence. Therefore, in the end of this process one mel-cepstral vector sequence $\mathbf{C} = \{\vec{c}_1, \ldots, \vec{c}_T\}$ and another pitch sequence $\mathbf{P} = \{P_1, \ldots, P_T\}$ are generated. Each component from each sequence (vector for mel-cepstral coefficients, and scalar for pitch) is repeated d_{q_j} times, for a given state q_j of the referred HMM.

2.3 Speech Synthesis from HMMs

The process of synthesizing speech consists on using the parameters generated from the HMM sentence as the input for the MLSA (Mel Log Spectrum Approximation) filter. The MLSA inputs mel-cepstral coefficients and pitch information and outputs the synthesized speech [4].

3 Experiments

In this section, five database utterances are analyzed with respect to pitch, spectrum and duration. The comparative tests can be used to define which characteristics of the speech synthesizer can degrade at most the overall intelligibility or naturalness.

3.1 Mode descriptions

Each one of the five sentences was synthesized by the proposed TTS according to seven different modes, namely:

- Mode 1 - original pitch, spectrum and durations;
- Mode 2 - synthesized spectrum with original pitch and durations;
- Mode 3 - synthesized spectrum and pitch with original durations;
- Mode 4 - synthesized spectrum and durations with original pitch;
- Mode 5 - synthesized pitch with original spectrum and durations;
- Mode 6 - synthesized pitch and durations with original spectrum;
- Mode 7 - synthesized spectrum, pitch and durations.

Table 1 shows an overview of these seven modes. It is important to observe that all the combinations among pitch, duration and spectrum were tested, except for original pitch and spectrum with synthesized duration, which does not make sense, since the original pitch and spectrum already include state duration informations.

The original state durations of those five sentences, which are segmented and labeled, was obtained after the HMM training of each of them separately, by the Viterbi algorithm. Original spectrum and pitch were generated through mel-cepstral analysis and autocorrelation, respectively.

Table 1: Systematic view of the tests, where 'O' means original and 'S' synthesized.

	Spectrum		Pitch		State Duration	
Mode	S	O	S	O	S	O
1		X		X		X
2	X			X		X
3	X		X			X
4	X			X	X	
5		X	X			X
6		X	X		X	
7	X		X		X	

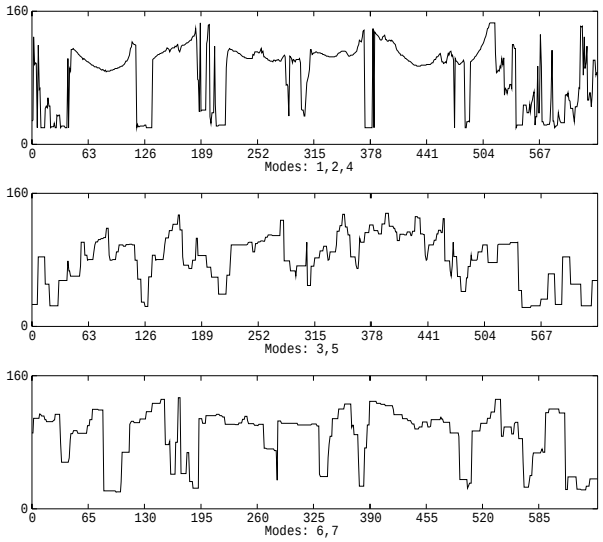

Figure 2: Pitch patterns according to the different modes: (a)Modes: 1,2,4; (b)Modes: 3,5; (c)Modes: 6,7.

3.2 Graphical Analysis

Figure 2 and Figure 3 refers to two different utterances from our speech database. The former figure refers to the Portuguese utterance: "O analfabetismo é a vergonha do país" which means "The illiteracy is the shame of the country", while the latter refers to: "Os maiores picos da Terra ficam de baixo d'água" which means "The highest mountains in the world are submerged."

Through the analysis of Figure 3(a) it can be noticed that with the original state duration the main acoustic events are synchronized, different from Figure 3(c), where duration models are estimated. Since Figure 3(b) is more similar to Figure 3(a) than Figure 3(c), we expect that Mode 2 will give better results than Mode 4. Evaluating Figure 2 we obtain a similar conclusion. Consequently, it is expected that Mode 5 gives better results than Mode 6.

3.3 Subjective Tests

A listening test was performed with 18 listeners. They gave their respective scores ranging from 0 to 5 concerning the quality degree for each one of the five sentences, where the reference corresponded to those sentences synthesized according to Mode 1. For this case, all the synthetic sentences were considered to have score 5. Table 2 shows the results obtained by this test. It can be noticed that these results are in accordance with the graphics analysis in Section 3.2.

Table 2: Results for the test where 18 listeners gave scores ranging from 0 to 5, related to the quality of the synthetic speech (Mode 1 = 5).

Mode	Average score	(%)
1	5.0	100.0
2	3.63	72.66
3	2.77	55.52
4	1.67	33.47
5	3.67	73.45
6	2.74	54.78
7	1.62	32.40

Analyzing modes 2 and 5, we can observe that both, spectrum and pitch information, have a similar impact on the quality of the TTS system.

According to modes 3 and 6, we can infer the difference between synthesizing spectrum and pitch with original duration, or modeling pitch and duration with original spectrum is imperceptible.

Even though mode 4, as modes 3 and 6, have two quantities synthesized and one original, it a lower score. We verified that the reason for this fact is that the duration modeling caused a mismatch between spectrum and

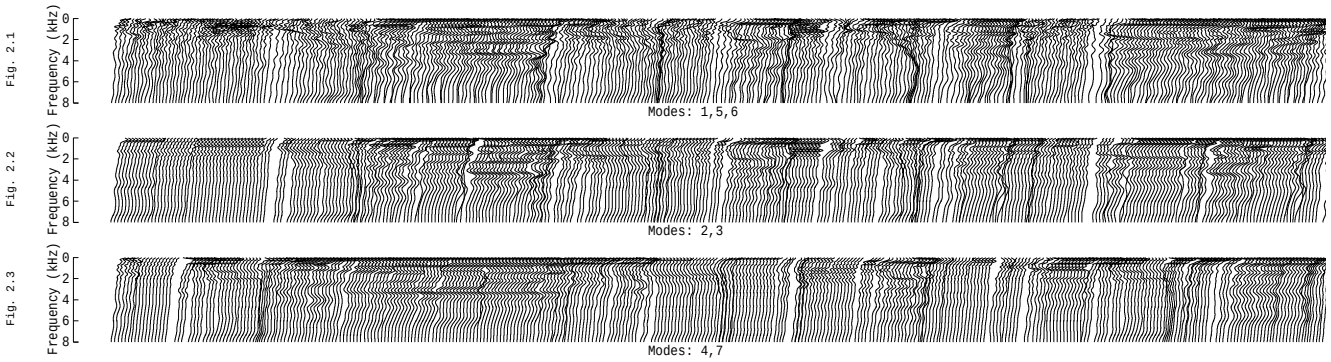

Figure 3: Running log spectrum of the mel-cepstral coefficients according to different modes: (a)Modes: 1,5,6; (b)Modes: 2,3; (c)Modes: 4,7.

pitch. The referred mismatch also caused the equal impact observed with modes 4 and 7.

4 Conclusions

In this work a Brazilian Portuguese TTS based on modeling of spectrum and pitch using HMMs is introduced. Since HMM is used to model the spectrum, speaker adaptation can be easily achieved. Experiments shows the effect of modeling pitch and spectrum using HMMs. Rule-based models for pitch and duration and comparison with the statistical models introduced in this work are subject of present research.

Acknowledgments

Special thanks to professor Keiichi Tokuda, for the useful discussions.

References

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP*, 1996.

[2] R. Donovan and P. Woodland, "Automatic speech synthesizer parameter estimation using HMMs," in *Proc. ICASSP*, pp. 640–643, 1995.

[3] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP*, pp. 1611–1614, 1997.

[4] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptative algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, pp. 137–140, 1992.

[5] G. O. Pinto, F. L. F. Barbosa, and F. G. V. Resende Jr., "A Brazilian Portuguese TTS based on HMMs," in *Proc. ITS*, 2002.

[6] J. Solewicz, "Text-to-speech synthesis for the Brazilian Portuguese (in Portuguese)," Master's thesis, PUC, Rio de Janeiro, RJ, Brazil, 1993.

[7] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, pp. 660–663, 1995.

[9] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. EUROSPEECH*, pp. 757–760, 1995.